Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR

Tülin ACAR, Hülya KELECİOĞLU***

Abstract

The aim of this research is to determine differential item functioning (DIF) by hierarchical linear modeling (HGLM) on test items and comparing these results by the DIF results determined by logistic regression (LR) and IRT-likelihood ratio (IRT-LR) techniques. Investigating the concordance between the techniques in determining the items with DIF, we have found significant relations between items with DIF determined in Turkish and Science sub-tests using LR and IRT-LR. In Social Studies test, we have found significant relation between the results of the HGLM and LR, HGLM and IRT-LR, LR and IRT-LR techniques. The number of items with DIF per gender determined by all three different techniques has been found to be almost half of the number of test items. Though all the level of the items determined by LR technique have been negligible, with IRT-LR technique only a very few of the items have been found with negligible DIF. Comparing the Social Studies and Science sub-tests with Turkish sub-test using HGLM technique, more than half of the items have been found to have DIF. Turkish sub-test has the maximum number of item with DIF.

Key Words

Differential Item Functioning, Hierarchical Generalized Linear Model, Logistic Regression, IRT-likelihood Ratio.

* *Correspondence:* Tülin Acar, PhD., Measurment and Evaluation Specialist, Parantez Education Research Publisher, Selanik Street, No:46/4, Kızılay-Çankaya-Ankara/Turkey.
E-mail:totbicer@gmail.com
** Assist Prof. Hülya Kelecioğlu, Hacettepe University, Department of Measurement and Evaluation 06800 Beytepe, Ankara/Turkey.

Kuram ve Uygulamada Eğitim Bilimleri / Educational Sciences: Theory & Practice
10 (2) • Spring 2010 • 639-649

The measurement of cognitive properties of students in various areas is examined by achievement tests or by ability tests in education. The aim of the competitive examinations is to choose the most fitted applicants from among different kinds of applicants. The selection of the up to grade applicants is related to the specifications of a qualified instrument of measurement. Whether being bias of the items that form the measuring instrument affects the properties of the measuring instrument. In the measurement results while examining the bias, Differential Item Functioning (DIF) is mostly utilized. DIF is the displaying differences of the probability of answering item correctly according to the subgroups in every ability level of psychological structure that is aimed to be measured with the item (Embretson & Reise, 2000; Lord, 1980). In DIF studies, the performances of different groups are compared according to the test items related to demographical specifications such as men-women in the same ability level, Asian-European etc. (Greer, 2004). Uniform-DIF is present, if the probability of answering an item correctly of the focused group is higher than the referred group in every ability levels. If the probability of answering an item correctly of the focused group differs from the referred group according to their ability levels, it is possible to talk about a non-uniform-DIF about the item (Zumbo, 2003).

The differential item functioning (DIF) is widely used for the research of bias in the measurement results. There are a lot of methods to determine DIF. Some of these methods are based on classical test theory (CTT). Mantel Haenzel (MH) technique which is largely used, logistic regression (LR) method and simultaneous bias test (subtest) method can be given as examples of the methods based on CTT (Gierl, Khalia, & Baughton, 1999). Some DIF determination methods are based on item response theory (IRT) and Lord's chi-square test, Raju's area measures and likelihood ratio are examples of this method (Camili & Shepard, 1994; Öğretmen, 1995).

DIF is the differentiation between the item and the probabilities of correct response to the item in every talent level of the psychological structure that will be measured (Embretson & Reise, 2000; Lord, 1980). There are many studies done to determine the DIF by Mantel Haenszel (MH) method, and by the developing methods, the number of DIF determination studies done by IRT-likelihood ratio and LR instead of MH have increased. However it has been observed that the data in

educational researches have a hierarchical structure and this situation have drawn attention on hierarchical generalized linear modeling (HGLM) in DIF determination studies (Atar, 2007; Subedi, 2005; Willms, 1999).

Hierarchical Linear Model (HLM) Method

Hierarchical linear model (HLM) provides a statistical model covering multilevel models (Greer, 2004; National Assessment of Educational Progress [NAEP], 2006). In groups researched level-1 represents the individual level while level-2 represents the group level. Considering that each group has different regression lines, it is easy to model mixed variables having multiple features and multiple group variables with different observation numbers by HLM (Gokiert & Ricker, 2004). Hierarchical linear models have been designed to sustain the assumption of the independency of the observations from each other for the tests where the individuals and the groups they belong to are tested together (Osborne, 2000; Raudenbush & Bryk, 1987). If the result variable is the measurement results on the alignment and classification level HGLM, a special form of HLM, is employed.

DIF Determining with Hierarchical Generalized Linear Model (HGLM)

If the outcome variable is measuring results in ordering or classification, HGLM can be used which is a special form of HLM. Thus, there is no necessity for a conversion process in the outcome variable. In the outcome variables having 2 categories, binom distribution is taken into account which is known as Bernoulli distribution and lojit connection function is used (Raudenbush & Bryk, 1986). The lojit connection function which is used for the binary outcome variable is used in this way:

$$\eta_{ij} = \log\left(\frac{\varphi_{ij}}{1-\varphi_{ij}}\right)$$

φ_{ij} in the equation is showing the probability of "to be" of the outcome variable and the outcome variable takes the values between 0 and 1. η_{ij} , is the logarithm of probability of "to be" (log-odds).

Predictive variables are added to model level-2 that are reflecting the specifications of the student -this is the DIF determining performance on the item- when it is needed to examine whether the student specifications have impacts on answering the test items correctly or not. In HGLM, Level-1 and level-2 equations that will be established in order to determine DIF with conditional modeling is as seen below (Williams, 2003):

Level-1 Equation (item level): in order to show the i (i=1,2,….k) item and j (j=1,2,….N) individual index

$$\eta_{ij} = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + ... + \beta_{(k-1)j}X_{(k-1)ij} + r_{ij}$$

η_{ij}: Estimated outcome variable, in other words, the probability of individual j. to give the correct answer to item i.

$X_{q\,ij}$: is indicator variable for item i. When the answer given to an item is on item i. ($q=i$), the value is 1 , in another condition ($q \neq i$) the value is 0.

β_{0j} : it is the breakpoint. When the all $X_{q\,ij}$'s become 0, the affect of the item which is not taken for the model occurs. For this reason, β_{0j}, is the effect of the item which is not taken for the model.

β_{1j} : is the effect of the 1. item on the probability (outcome variable) of individual j. to give the correct answer up to i=1,2,…(k-1). Parameters from β_{1j} to $\beta_{(k-1)j}$ is a coefficient that shows the effects of the items on the probabilities of giving the correct answer for the individual from 1. item to item k. Individual j. is associated with different individuals and different item level parameters. If the level increases, j. in β_{ij} decreases and the item parameters kept instant between individuals.

2. Level is formed in order to see the differences between the probabilities of answering each item correctly according to the genders of the students (Williams, 2003).

Level 2 (student level) Equation:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(gender)_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(gender)_j$$
…
$$\beta_{(k-1)j} = \gamma_{(k-1)0} + \gamma_{(k-1)1}(gender)_j$$

β_{ij}: is the effect of item i. on the probability of giving the correct answer for individual j. up to i=1,2,...(k-1). Parameters from β_{1j} to $\beta_{(k-1)j}$ are the effects of the items on the probability of giving the correct answer form the 1. item to item k. for individual j.

γ_{00}: is the referred item parameter.

γ_{01}: is the effect of the probability of correct answering of item i. on gender variable.

u_{0j}**:** Effect of random gender variable. It is the random effect of β_{0j} which shows normal distribution that has distribution average 0 and variance τ.

DIF Determining with Logistic Regression (LR)

LR, is a kind of regression that can be applied when dependant variable isn't continuous variable and it is a special regression model where dependant variable (item scores) may get a dicategorious result rate (Jodoin & Gierl, 2001). If the performances of the group members on an item are estimated with logistic regression method, it is possible to talk about a DIF on that item (Swaminathan & Rogers, 1990). For this reason, LR is a method at the same time which is used in order to find out the items containing DIF. With LR method it is possible to determine both uniform-DIF and non-uniform-DIF. The level of effect can be determined as well. To do this, the standardized regression parameters can be used. Jodoin and Gierl (2001) are classifying the effect levels of DIF that are determined with LR in this way.

A Level: if $R^2 < 0.035$, a negligible level of DIF is present.

B Level: if $0.036 < R^2 < 0.070$ a medium level of DIF is present.

C Level: if $R^2 > 0.071$ a magnitude level of DIF is present.

DIF Determining with Item Response Theory- Likelihood Ratio (IRT-LR)

IRT provides the facility of obtaining item scaling independent of individuals and ability parameters independent of items, which cannot be explained by CTT, through series of mathematical models. In other words, it is claimed that the ability estimation in DIF are made independent of both chosen items and the performance of the people

taking the test. According to IRT (Hambleton, Swaminathon, & Rogers, 1991),

(i) The probability of answering on item correctly is independent of the ability level of the answering group, who answers that item,

(ii) The ability of an individual is independent of any item group that is applied to him.

(iii) It is possible to predict the features of a test before the application of the test.

IRT is quite a practical method with its mathematical theoretical structure and is more information than CTT. As IRT is a model which can define the relation among the answers of individuals, it can be determined in DIF through this method (Greer, 2004; Kim, 2003). A strong part of DIF determining with Item Response Theory (IRT) is the utility of item response curves and item characteristic curves (Thissen, 2001). If, an item functions different in focused groups and referred groups, in other words, if the item response curves are different for two groups, presence of DIF is applicable. For both groups, the item parameters are estimated and the estimated item parameters are compared according to DIF with IRT method. Many software have been developed with IRT-LR technique, in determining DIF. In research, results of IRTLRDIF software was used in DIF determining with IRT-LR technique.

In determining of DIF with Likelihood Ratio, IRTLRDIF program, which was more practical than Multilog program, was setup by Thissen (2001). The hypothesis of absence, which is built while analyzing DIF determining with Likelihood Ratio, is as "there is no significant difference between the item parameters that are calculated from focused and referred groups". In IRTLRDIF program, the results of the compact model (CM) for the test of absence hypothesis and the augmented model (AM) are compared. In the compact model, the parameters of all items in focused and referred groups are supposed to be equal, in other words, none of the items are assumed as DIF. In the augmented model, it is supposed that parameters of item i. for the focused and referred group can differ, and for the other items the parameters are supposed to be equal as happened in the augmented model. While a likelihood function can be obtained from compact model, as many likelihood functions as the number of items can be obtained from the augmented model. G^2

value is obtained by taking the logarithms of the likelihood function of the compact model and the augmented model (Thissen, 2001).

$$G^2 = -2LL_C - (-2LL_A)$$

G^2 is showing the chi square distribution. Number of the item parameters is the degree of independence of the distribution. In the condition of value G^2 exceeding value 3.84 ($G^2_{sd=1;\alpha=0,05}$) the null hypothesis is denied and the presence of DIF is possible for the related item (Thissen, 2001). The quantitative value of G^2 value appoints the effect degree of DIF. Taking into account Cohen's G^2 statistics, the classification made for the degree of effect is as seen below (Greer, 2004):

A Level, if $3.84 < G^2 < 9.4$ a negligible level of DIF is present.

B Level, if $9.4 < G^2 < 41.9$ a medium level of DIF is present.

C Level, if $G^2 > 41.9$ a magnitude level of DIF is present.

Purpose of the Study

The aim of this research is to determine DIF by HGLM on test items and comparing these results by the DIF results determined by LR and IRT-LR techniques. As it is a frequent case to have nested data in education, it is thought to be significant to evaluate the DIF determination phases using HGLM and to compare the DIF determination results obtained using HGLM with other methods. Accordingly in this research, we have first determined whether there is DIF in the items of Turkish, Social Studies and Science sub-tests of 2006 Secondary Education Institutional Examination (SEIE) using the HGLM, LR and IRT-LR techniques. We investigated whether there is a concordance between the items having DIF according to the sub-tests determined with these methods.

Method

Sample

As the DIF determination study is conducted per gender in this research, we have formed subgroups as gender variable. 5423 (50.6%) of the sampling students are females while 5304 (49.4%) are males. The focus group is the female group and the reference group is the male

one. We used HLM-6.04 for HGLM (Raudenbush, Bryk, Cheong, & Congdon, 2001), the script prepared in Zumbo's SPSS for LR analysis and IRTLRDIF for IRT-LR.

Instrument

In this research, 2006 SSIE results have been used as data in order to inspect the DIF determining techniques. For this reason, there is no interpretation concerning the contents of the items that show DIF. SSIE is consisted of 25 itemed Turkish, Social Sciences, Maths and Science sub-tests. It has been designated that, the reliability coefficient of Maths sub-test was (α=0.688) low; according to the factor analysis technique the test was not single-dimensional and G^2 values designated with IRT-LR technique have taken excessive values. For this reason, Maths test was exempted from the analysis. Turkish 0.849, Social Sciences 0.873 and Science 0.792 were the Cronbach alpha reliability coefficient.

Data Analysis

The data were analyzed with HGLM, LR, IRT-LR techniques. For HGLM analysis HLM-6.04 (Raudenbush et al., 2001), for LR analysis script which was prepared in Zumbo's SPSS program (Zumbo, 1999) and for IRT-LR analysis IRTLRDIF (Thissen, 2001) programs have been used.

Results

In Turkish, Social Studies and Science sub-tests we have respectively determined items with DIF using the following techniques; 20, 11 and 14 items with HGLM, 21, 16 and 18 items with LR, 17, 16 and 19 items with IRT-LR.

Investigating the concordance between the techniques in determining the items with DIF, we have found significant relations between items with DIF determined in Turkish and Science sub-tests using LR and IRT-LR. In Social Studies test, we have found significant relation between the results of the HGLM and LR, HGLM and IRT-LR, LR and IRT-LR techniques.

The number of items with DIF per gender determined by all three different techniques have been found to be almost half of the number of test items. Though all the level of the items determined by LR technique have been negligible, with IRT-LR technique only a very few of the items have been found with negligible DIF. Comparing the Social Studies and Science sub-tests with Turkish sub-test using HGLM technique, more than half of the items have been found to have DIF. Turkish sub-test has the maximum number of item with DIF. Öğretmen (2006), Yıldırım (2006), Roever (2005), Shen (1999), Takala & Kaftandjieva (2000) studied with this method and they found DIF in test items. They also made some suggestions regarding their findings.

It is possible to increase the quality of the questions in question banks by determining which technique is best for which test to determine DIF by producing data, using simulation studies, similar to the parameters of the selection and placement examinations like SEIE. It is possible to conduct DIF investigations on test items by forming up different subgroups like DIF analysis, socio-economical level and school type.

It is necessary to conduct DIF investigation of large scale placement examination every year. Additionally, the items with DIF should be investigated for bias and the tests should be regulated in accordance with the results of these investigations.

In this research, we have made DIF analysis on test items using HGLM, IRT-LR and LR techniques. It is possible to conduct DIF investigations on other test items and DIF analysis techniques which have not been included under the scope of this research. It is also possible to compare analysis results conducted by using DIF techniques in different sampling sizes.

References/Kaynakça

Atar, B. (2007). *Differential item functioning analyses for mixed response data using irt likelıhood-ratio test, logistic regression, and GLLAMM procedures.* Yayınlanmamış doktora tezi, Florida State Üniversitesi.

Bryk, A. S. & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change, *Psychological Bulletin*, *101*(1), 147-158. Erişim: 10 Şubat 2008, http://www.personal.psu.edu/jxb14/M554/articles/Bryk&Raudenbush1987.pdf.

Camili, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test items* (Vol. 4). California: SAGE Publications.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. USA: Lawrence Erlbaum Associates, Mahwah, NJ.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999). *Gender differential item functioning in mathematics and science: Prevalence and policy implications.* Paper presented at the Symposium Entitled "Improving Large-Scale Assessment in Education" at the Annual Meeting of the Canadian Society for the Study of Education, Canada.

Gokiert, R. J. & Ricker, K. L. (2004). Gender differential item functioning on the WISC-II: Analysis of the Canadian standardization sample. *Centre for Research in Applied Measurement and Evaluation*. Erişim: 24 Mart 2010, http://www2.education.ualberta.ca/educ/psych/crame/files/Gender_DIF_WISCIII.pdf

Greer, T. G. (2004). *Detection of differential item functioning (DIF) on the SATV: A comparison of four methods: Mantel-Haenszel, logistic regression, simultaneous item bias and likelihood ratio test.* Unpublished doctoral dissertation, University of Houston.

Hambleton, R. K., H. Swaminathan & J. H. Rogers. (1991) *Fundamentals of Item Response Theory,* Sage Publications, Boston.

Jodoin, G. M. & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Kim, W. (2003). *Development of a differential item functioning (DIF) procedure using the hierarchical generalized linear model: A comparison study with logistic regression procedure.* Yayınlanmamış doktora tezi, The Pennsylvania State Üniversitesi.

Lord, M. F. (1980). *Applications of item response theory to pratical testing problems*. USA: Lawrence Erlbaum Associates, Broadway, Hillsdile, NJ.

Millî Eğitim Bakanlığı. (2008). *Ortaöğretime geçiş sistemi Seviye Belirleme Sınavı ve 6 ve 7'nci sınıflar Devlet Parasız Yatılılık ve Bursluluk Sınavı e-kılavuz*. http://oges.meb.gov.tr/sbs/docs/sbs_kilavuz.pdf adresinden 5 Ocak 2009 tarihinde edinilmiştir.

National Assessment of Educational Progress. (2006). *Comparing private schools and public schools using hierarchical linear modeling*. Retrieved February 10, 2008, from http://www.nces.ed.gov/NAEP/pdf/studies/2006461.pdf.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. practical assessment. *Research and Evaluation, 7*(1). Retrieved February 14, 2007, from http://www4.ncsu.edu/~jwosbor2/otherfiles/MyResearch/PARE2000-HLM.pdf.

Öğretmen, T. (1995). *Differential item functioning analysis of the verbal ability section of the first stage of the university entrance examination in Turkey*. Yayımlanmamış yüksek lisans tezi, Orta Doğu Teknik Üniversitesi, Eğitim Bilimleri Enstitü, Ankara.

Öğretmen, T. (2006). *Uluslararası Okuma Becerilerinde Gelişim Projesi (PIRLS) 2001 Testinin psikometrik özelliklerinin incelenmesi: Türkiye-Amerika Birleşik Devletleri örneği*. Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Raudenbush, S. & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59(1), 1-17.

Raudenbush, S. W., Bryk, A. S. Cheong, Y. F., & Congdon, R. T. (2001). *HLM 5 hierarchical linear and nonlinear modelling*. Lincolnwood, IL: Scientific Software International.

Roever, C. (2005). *That's not fair! Fairness, bias and differential item functioning in language testing.* Erişim: 7 Şubat 2008, http://www2.hawaii.edu/~roever/brownbag.pdf.

Shen, L. (1999). *A multilevel assessment of differential item functioning*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada Retrieved 20 June 2009 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/17/97/c8.pdf.

Subedi, B. R. (2005). *Demonstratıon of the three-level hierarchical generalized linear model applied to educational research.* Unpublished doctoral dissertation, The Florida State University.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Takala, S. & Kaftandjieva, F. (2000). Test Fairness: A DIF analysis of an l2 vocabulary test. *Language Testing*, *17*(3), 323-340.

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics ınvolved in item response theory likelihood-ratio tests for differential item functioning.* Retrieved February 10, 2008, from http://www.unc.edu/~dthissen/dl.html.

Willms, J. D. (1999). Basic concepts in hierarchical linear modeling with applications for policy analysis. In G. Cizek (Ed.), *Handbook of educational policy* (pp. 473-493) San Diego, CA: Academic Press. Retrieved February 10, 2008, from http://www.unb.ca/crisp/pdf/9806land.pdf.

Williams, N. J. (2003). *Item and person parameter estimation using hierarchical generalized linear models and polytomous item response theory models*. Unpublished doctoral dissertation, The University of Texas at Austin.

Yıldırım, H. H. (2006). *The differential item functioning (DIF) analysıs of mathematics items in the international assessment programs.* Unpublished PHD doctoral dissertation, Middle East Technical University, Eğitim Bilimleri Enstitüsü, Ankara.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Retrieved February 10, 2008, from http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*(2), 136-147.